

# Probabilistic paths in protein interaction networks

Hailiang Huang<sup>1,2,3</sup>, Lan V. Zhang<sup>4</sup>, Frederick P. Roth<sup>4</sup>, and Joel S. Bader<sup>1,2</sup>

**Keywords:** protein physical interaction, Monte Carlo simulation, greedy algorithm

Understanding how proteins are physically organized into complexes and pathways is increasingly based on observations from high-throughput experiments. Yeast is a widely used model for eukaryotic proteomics. Evidence from high-throughput experiments is unreliable, however, due to high false-positive and false-negative rates. We investigate algorithms for inferring protein complexes from noisy proteomics data, expanding a subset of known members into a full complex. The protein interaction data is represented as an undirected weighted graph, with proteins as vertices and edge weights representing the confidence measures. We use confidence measures estimated previously by naïve Bayes (NB) [1], logistic regression (LR) [3], and decision trees (DT) [4].

We investigate two classes of algorithms, deterministic and probabilistic. The deterministic algorithms, BESTPATH [2], SPE [1], and SUMPATH (this work), calculate the threshold neighborhood around each seed protein, with the threshold depending on edge weights rather than on the raw number of links. The probabilistic algorithms, PRONET [1], PROPATH-EXP (this work), and PROPATH-ALG (this work), generate an ensemble of networks using the edge weights as probabilities that each edge occurs.

We compare the performance of these algorithms by assessing their ability to extract a known complex based on partial knowledge of its components. We use 23 known complexes from MIPS and generate 10 random 50-50 splits of the complex into seed proteins and target proteins. The seeds are used as input seeds for each of the algorithms, which returned lists of proteins ranked by decreasing likelihood of membership in the same complex as the seeds. This ranked list is used to calculate TP (true-positive) rate and FP (false-positive) rate by comparing with the target protein list. Performance is visualized by graphing the TP rate vs. FP rate graph (Fig 1). Quantitative measures such as normalized AUC (Area Under the Curve) and FP-50 (false-positive rate at 50% recall) help to rank the algorithms (Table 1). A graph depicting the recovered complex (Fig 2) offers a direct visualization.

We find that BESTPATH, a simple deterministic algorithm, performs at least as well as the other algorithms, including the novel probabilistic algorithms.

The PRONET algorithm was designed for edge weights defined as the probability of a direct interaction, which is how the NB confidence scores were trained. The other algorithms permit edge weights to also include the probability that proteins are co-complexed with or without a direct interaction, which is how the DT confidence scores were trained. The LR scores were trained similarly to the DT scores, except that the prior estimate for  $\text{Pr}(\text{edge})/\text{Pr}(\text{no edge})$  was set to 1 rather than fit or optimized. Thus, it is not strictly appropriate to compare results for PRONET with the LR or DT networks. Nevertheless, even for the NB network, the BESTPATH algorithm

---

<sup>1</sup> Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, E-mail: hlhuang@pha.jhu.edu

<sup>2</sup> High-Throughput Biology Center, Johns Hopkins School of Medicine, Baltimore, MD 21287

<sup>3</sup> Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218

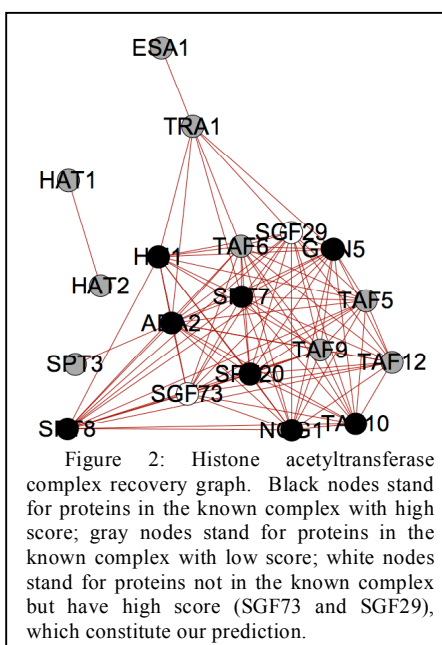
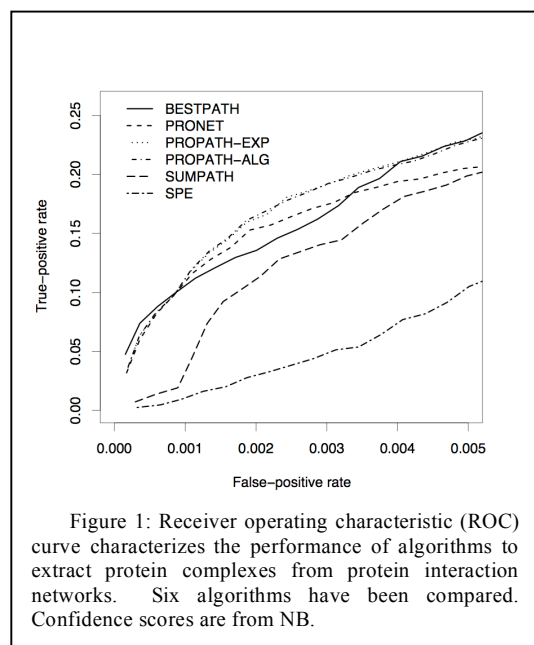
<sup>4</sup> Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115

provides better results for this limited test. The other probabilistic algorithms may be more robust to errors in the confidence score statistical models.

## Figures and tables

Type of algorithm	Algorithms	Ref	Avg. Rank (1 = best) / NB Rank	AUC 0.5%(%) <sup>a</sup>			FP-50(%) <sup>b</sup>		
				NB	LR	DT	NB	LR	DT
Deterministic	BESTPATH	[2]	1.33 / 2	16.9 <sup>3</sup>	36.6 <sup>1</sup>	35.3 <sup>1</sup>	7.69 <sup>1</sup>	0.9 <sup>1</sup>	0.9 <sup>1</sup>
Deterministic	SUMPATH	This work	3.66 / 3.5	13.2 <sup>5</sup>	0.113 <sup>6</sup>	9.18 <sup>2</sup>	8.08 <sup>2</sup>	21.6 <sup>5</sup>	5.84 <sup>2</sup>
Deterministic	SPE	[1]	4 / 4.5	5.35 <sup>6</sup>	3.2 <sup>4</sup>	1.74 <sup>3</sup>	10.5 <sup>3</sup>	5.8 <sup>4</sup>	9.28 <sup>4</sup>
Probabilistic	PROPATH-EXP	This work	2.83 / 2.5	18.1 <sup>1</sup>	33 <sup>2</sup>	1.74 <sup>3</sup>	31.6 <sup>4</sup>	1.36 <sup>2</sup>	9.35 <sup>5</sup>
Probabilistic	PROPATH-ALG	This work	2.83 / 2.5	18.1 <sup>1</sup>	32.8 <sup>3</sup>	1.74 <sup>3</sup>	31.6 <sup>4</sup>	1.37 <sup>3</sup>	9.27 <sup>3</sup>
Probabilistic	PRONET	[1]	5.17 / 4	16.7 <sup>4</sup>	0.264 <sup>5</sup>	0.6 <sup>6</sup>	31.6 <sup>4</sup>	29.8 <sup>6</sup>	48.4 <sup>6</sup>

Table 1: Summary of methods. <sup>a</sup> AUC 0.5%: area under the curve (AUC) at a false-positive rate of 0.5%, in percentage scale. For each network, each algorithm was ranked 1-6 in performance, 1 = best, 6 = worst, for AUC-0.5 and FP-50. The ranks were averaged to give an overall measure of each algorithm's performance. The ranks for the NB network, which contributed to the average, are shown separately.



## References

1. Asthana S, King OD, Gibbons FD, Roth FP: Predicting Protein Complex Membership Using Probabilistic Network Reliability. *Genome Res* 2004, 14(6):1170-1175.
2. Bader JS: Greedily building protein networks with confidence. *Bioinformatics* 2003, 19(15):1869-1874.
3. Bader JS, Chaudhuri A, Rothberg JM, Chant J: Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* 2004, 22(1):78-85.
4. Zhang LV, Wong SL, King OD, Roth FP: Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics* 2004, 5(1):38.